informationresearch

VOL. 15 NO. 4 DECEMBER, 2010

Contents | Author index | Subject index | Search | Home

Analysis of automatic translation of questions for question-answering systems

Lola García-Santiago and *María-Dolores Olvera-Lobo*
CSIC, Unidad Asociada Grupo SCImago, Madrid, España.
University of Granada, Department of Library and Information Science, Colegio Máximo de Cartuja s/n., 18071 Granada, Spain

Abstract

Introduction. *Multilingual question-answering systems can provide users with specific data in response to queries by searching for a minimal fragment of text that applies to the query, regardless of the language in which the question is formulated and the answer is found. The aim of this paper is to analyse the automatic translation of questions (intended as queries input to a cross-language, question-answering system) from German and French into the Spanish language.*
Method. *The methodology used for evaluation, based on automatic and subjective measures, appraises whether the translation will serve as input to a system. That is, does the question retain its validity and fulfil its function, allowing a proper response to be found?*
Analysis. *The main features of multilingual question-answering systems are described and then we analyse the effectiveness of the translations achieved through three popular online translating tools: Google Translator, Promt and Worldlingo.*
Results. *Our findings serve to identify which is the most reliable translator for both pairs of languages overall. However, an even more reliable option would be to use two different translators, depending on which of the two source languages is being dealt with.*
Conclusions. *The results contribute to the realm of innovative search systems by enhancing our understanding of online translators and their potential in the context of multilingual information retrieval.*

CHANGE FONT

Introduction

Information retrieval is the collection of tasks implemented by the user to locate and access the information sources that are appropriate to resolve a given need for information. In these tasks, indexing languages, abstracting techniques, and the description of the documental object play key roles, largely determining how fast and efficient retrieval is (Belkin and Croft 1987). Ideally, there is a balance between the precision and the recall of the informational yield. This aspect is increasingly important, as the World Wide Web diffuses vast quantities of new content every day, in a great variety of formats and languages (Turpin and Hersh 2001).

When a need for information arises, a process called the search strategy is set in motion, which leads to the supply of documents by the system (Belkin and Croft 1987). This process entails seven basic stages:

a. definition of the information need;
b. selection of the information sources to be used;
c. translation of the user query expressed in natural language into the indexing language of the information source, if necessary;
d. translation of the expression from the indexing language to the query language of each information system;
e. implementation of expressions obtained from the query language;
f. assessment of results by the user and the redefinition of the query expressions if the results are not relevant; and
g. selecting and obtaining the documents that respond to the user’s needs.

One step in the evolution toward improved retrieval resides in the use of question-answering systems, which pursue the supply of specific data instead of documents and respond to the questions formulated by users in natural language (Hallet *et al.* 2007). If this answer derives from documents that are found in other languages, the situation involves a translingual or multilingual question-answering system. This type of system is particularly complex, as it incorporates the capacities of a translingual system for cross-language information retrieval, while also working as a question-answering system.

This approach, where access entails gathering accurate data that respond to specific questions, is a specialization within the more generic concept of multilingual access to information (López-Ostenero *et al.* 2004). In important international forums, principally the Text REtrieval Conference (Voorhees and Harman 2000) and the Cross-Language Evaluation Forum (Clough *et al.* 2004, 2006), such systems have been evaluated, with discussion of innovative proposals and techniques. Both have set up a track dedicated to studying question-answering systems. In the general scope of cross-lingual information retrieval, there are several proposals intended to overcome the language barrier that appears when queries and the documents obtained are in different languages.

As mentioned above, systems that deal with multiple languages usually rely on a translation module. The architecture of a cross-lingual information-retrieval system would use one of three main approaches: the translation of the query, the translation of documents from the database, or an interlinguistic approach (Oard and Diekema 1998). However, at present, translation of the answer could be another possibility (Bos and Nissim 2006). Translating the query is the most frequent option since they are shorter texts than the documents, and therefore their translations have limited computational costs (Hull and Grefenstette 1996). Nevertheless, many researchers describe the problems that arise in the translation process when the questions are short and offer little context to help eliminate any semantic ambiguity in the terms of the question; in such cases, interaction with the user (Oard *et al.* 2008) may improve the results. The underlying translation processes apply different linguistic resources, such as bilingual dictionaries, textual corpora, machine-translation software, and thesauri (López-Ostenero *et al.* 2004; Abusalah *et al.* 2005); and various mechanisms for disambiguation and the selection of the most appropriate translation between the different proposals available (Kishida 2005).

As depicted in Figure 1, for the systems that incorporate a translation module, the user enters a specific query, generally including some interrogative adverb (e.g. How? When? Where?) in a given natural source language. This question is translated by an automatic translator. The resulting search expression will then be the *input*, or the formulation of the query to be used by the search engine of the system for comparing and matching it with the documents in the database. Once documents relevant to the query are located, the system breaks them up into sections, selects the excerpts that include the candidate responses and selects a final response. This response, along with its location in the corresponding document, is finally delivered to the user.

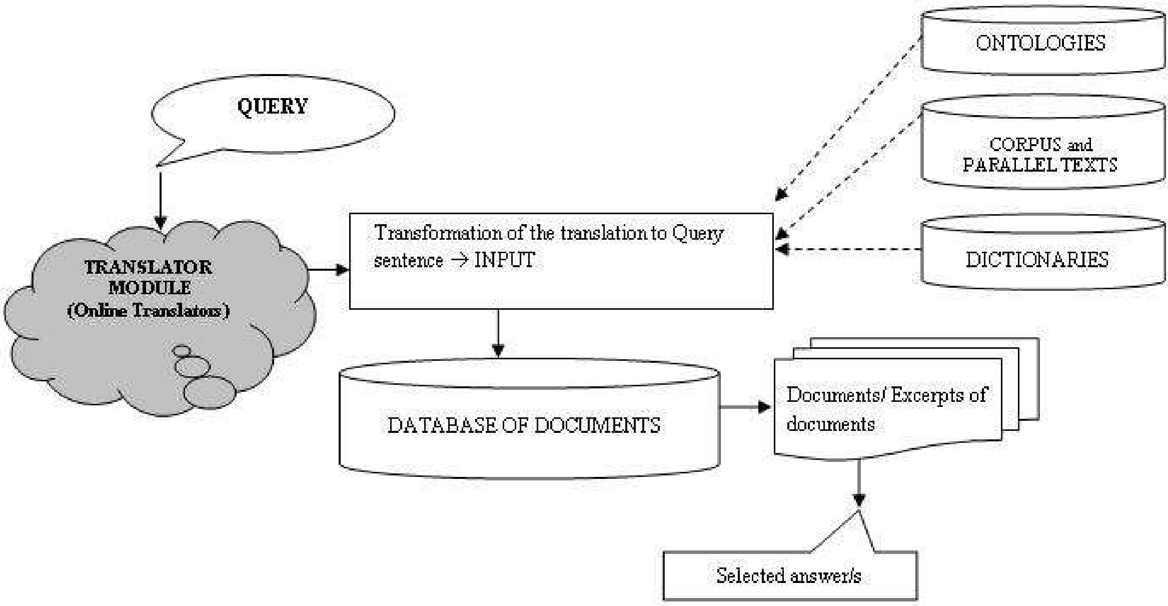

QUERY
TRANSLATOR MODULE (Online Translators)
Transformation of the translation to Query sentence → INPUT
ONTOLOGIES
CORPUS and PARALLEL TEXTS
DICTIONARIES
DATABASE OF DOCUMENTS
Documents/ Excerpts of documents
Selected answer/s

Figure 1: Basic elements of a translingual question-answering system

Given this background, our study focuses on the first module of the translingual question-answering systems, designed to translate the original user query. In the following sections, we present a comparative study of the quality of the different automatic translation tools that may be used online for no charge, applying three that translate from German and French into the Spanish language. Our perspective is a documental one; that is, we analyse the functionality of the translator as a mediating instrument in the search for answers. To this end, we apply well-known assessment measures (both objective and subjective ones) for machine translation, with the help of EvalTrans software. Finally, we analyse the results and offer succinct conclusions.

Evaluation of automatic translations as input in cross-language question-answering systems

One of the objectives of this study is to identify the most adequate online translator for a given question-answering system entailing a collection of documents in Spanish. In this case, the questions would be formulated in French or in German and would have to be translated into Spanish in order to constitute the system input before proceeding.

Our study used a set of questions from the Cross-Language Evaluation Forum, 2008, with 200 queries in German and the same 200 in French. The questions as expressed in Spanish by each of the online translators were both manually and automatically analysed, applying objective and subjective measures for the evaluation of automatic translation with the aid of EvalTrans software.

Online translators evaluated

Google Translator, *Promt* and *Worldlingo* were selected for this study because they allow us to translate and compare results using language pairs of German to Spanish and French to Spanish. Moreover, they are widely used services, they are quick in translating and they show reasonable quality overall. There are limitations regarding the maximum amount of text (from 150 to 300 characters) with which the free online translators can work, except *Google Translator*, which admits much more extensive texts, but these do not interfere with the purposes of our study, as a question-answering system deals with specific questions for which the formulation is not excessively long.

Google Translator is an automatic translator; that is, it works without the intervention of human translators, using state-of-the-art technology instead. Most commercial machine-translation systems in use today have been developed using a rule-based approach and require substantial work to define vocabularies and grammars. However, this system takes a different approach; the computer is fed billions of words of text, both monolingual text in the target language as well as aligned text consisting of examples of human translations between the languages. Afterwards statistical learning techniques are applied to build a translation model which is able to contextualise. *Promt* uses semantic classes to improve the syntactic and the semantic correspondence in a sentence.

A subdivision of translation systems into transfer-based systems and Interlingua-based systems has been adopted (Sokolova 2009). This subdivision is based on aspects of architectural solutions relating to linguistic algorithms. Translation algorithms for transfer-based systems are built as a composition of three processes: analysis of the input sentence in terms of source language structures, conversion of this structure into a similar target-language structure (*transfer*) and, finally, synthesis of the output sentence according to the constructed structure. Interlingua-based systems assume *a priori* that a certain structure metalanguage (*Interlingua*) is available, which, in principle, can be used for describing any structure of both the source and the target languages. *Wordlingo* uses statistical techniques from cryptography, applying machine-learning algorithms that automatically acquire statistical models from existing parallel collections of human translations. These models are more likely to be up to date, appropriate and idiomatic, because they are learned directly from real translations. The software can also be quickly customized to any subject area or style and do a full translation of previously unseen text. Statistical machine translation was once thought appropriate only for languages with very large amounts of pre-translated data. Additionally, with customization, such systems can also *learn* to translate highly technical material accurately (Grundwald 2009).

Sample and types of questions

In the stage of the query analysis, the question-answering system examines the user's question and determines what type of information is being requested. The classification of the questions is crucial for the system, as this information will be utilized in the search stage and in the selection and extraction of the potential responses (García Cumbreras *et al.* 2005). The collection of 200 Cross-Language

Evaluation Forum questions were formulated in German and in French in the input stage (before transformation to an indexing language) and were meant to gather precise data on a given subject.

A sample of 200 cases by each pair of the two languages, French and German, to Spanish is analysed. It is a sample that reflects a normal distribution for an infinite population. This sample has a 95% confidence level, a significance level of 2.5% (p = 0.025) and a statistical power of 0.8. Bloom, according his taxonomy, considers these queries *questions of knowledge* (Bloom 1956 and Forehand 2005). As in a survey, factual questions can be answered by either *Yes* or *No*. In other cases the answer is nominal; for example, a person, a place or institution. In addition, there are other knowledge questions. Closed-list questions imply a closed group of nominal possibilities (names, verbs, adjectives). Finally, there are definitional questions which describe the existence of some person, thing or process.

These questions may be classified as three types (Cross-Language... 2008):

- **Factual questions:** which seek tidbits of information such a names of person, organization or place,locations, time, quantities (measurement or recount) or object, among others. The question always starts with any interrogative particle: *Who...?, When...?, Where...?, Where...?,...*.
- **Definitional questions:** where the information solicited may entail synonymy and formulation with respect to a person, object, organization or concept (*What or who is it*?)
- **Closed list:** questions that call for a response enumerating various items (*What/is...? How...? or Name...*)

Unlike opinion questions that request an interpretation of the information, e.g. What should I do? or Is it better to do X or Y?, factual questions are focused on finding an specific and objective answer. According to Sloman (2005), these can cover from the most basic issue, which refers to a proposition that is capable of being true or false and requests the information whether the proposition is true or false, to issues which create one or more gaps in the proposition and specify requests for information. Some factual questions (Chan *et al.* 2003) can require information which is expressed by an interrogative particle, e.g., who (person), where (location), how many (number). However, for most questions using the interrogative word 'which' or what', we need to find the core noun to help us to identify the information required. For example: Which city is home of Superman? or What type of bee drills holes in wood? These kinds of questions are not lengthy and they have a formal and clear structure.

The 200 questions used were mostly (156) of the factual type; and in the Spanish language, they present mainly interrogative particles, as explained above. The remainder (44) are divided fairly equally into definition questions (24) and closed-list questions (20). Some examples are:

- *Où l'article de Gerda Taro sur cette bataille a-t-il été publié?*
- *Qu'est-ce que la vexillologie?*
- *Quels sont les pays membres de l'Observatoire européen austral?*/ Welche Länder sind Mitglied bei der Europäischen Südsternwarte?

Natural language is not an exact science. Interrogative adverbs in each language change their characteristics and require modifications when these subgroups must be identified clearly. For example, the question 'What political party does Tony Blair belong to?' differs from 'What is a screwdriver?' as the former contains a preposition which goes with the interrogative adverb. Therefore, the latter sentence is classified as a definition-type question. On the other hand, all kinds of questions can include a temporal restriction (e.g.. What book did George Orwell write in 1945?). Hermjakob (2001) classifies questions from a functional perspective during the information search on the Web. Furthermore, he identifies questions which include interrogative adverbs and he emphasizes questions that only ask for a confirmation or disconfirmation. This author sorts out this typology according to the answer rather the question;consequently, he tags them as 'Yes-No questions' and 'True-False questions'.

Measures for the evaluation of the automatic translation of questions

Evaluation of machine translation is an unresolved research problem that has been addressed by numerous studies in recent years (Ueffing and Ney 2007). The most extensively used assessment tools are classified into two major groups: automatic objective methods and subjective methods (Tomás *et al.* 2003). The objective evaluation methods compare a set of correct reference translations against the set of translations produced by the translation software under evaluation. The measurement units most often used work at the lexical level, comparing strings of text.

Our study evaluated the online translators in the light of the following parameters:

- *Word error rate* (WER) (Tillman *et al.* 1997; Vidal 1997): This is based on comparison of the Levensthein Distance or edit distance between two strings of characters. It measures the number of insertions, substitutions and eliminations necessary to convert one string into another (Figure 2). Yet, unlike the edit distance, which does this through the characters, *word error rate* calculates this distance in relation to the words involved. It can be seen as a *pessimistic* means of measurement for the evaluation of an automatic translation system (Pérez *et al.* 2004), in that, if the output of the system does not coincide exactly with the reference string, it is penalized, even when the expression is considered acceptable by a human translator.

$$WER[\%] = \frac{\sum_{i=1}^{n} d(t_i, t_i^r)}{\sum_{i=1}^{n} |t_i^r|} * 100$$

Figure 2: Word error rate formula

- aWER *All references word error rate* (aWER) (Tomás *et al.* 2003): which gathers, for comparison, all the reference phrases that may be taken into fair consideration, i.e., those that a human translator would include in the process of assessing translations. In other words, it corresponds to the rate of words which have to be changed, erased or included to achieve a correct translation. In this case, however, we have all the phrases of reference to compare, not only the first and, consequently, more alternatives.
- *Sentence error rate* (SER): which compares phrases (the string of output from the automatic translator and the reference string) overall, as units. It identifies or measures the lack of fit found through this comparison, which does not necessarily mean that the translation is erroneous. If the two strings differ in some way, the system is penalized. This may be considered an excessively tough tool (Pérez *et al.* 2004).
- *All references sentence error rate* (aSER) (Tomás, *et al.* 2003): the *sentence error rate* compares the phrase to be evaluated with a single reference string, whereas the *all references sentence error rate*, derived from the *all references word error rate*, is used in order to arrive at the percentage of phrases whose

translations are *incorrect*.

All the means of measurement described above are applied automatically. Therefore, the translations and the phrases of reference are compared without any specific determination of the type of error or discrepancy occurring between the two strings under consideration (Figure 3). For this purpose, there are other types of metrics requiring human intervention for evaluation. In the context of translingual question-answering systems that include the machine translation architecture, the aim of translation is more practical; so other evaluation measures of a subjective nature, such as *subjective sentence error rate* (sSER), were applied.

$$sSER(s_1^n, t_1^n)[\%] = \frac{100}{K \cdot n} \sum_{i=1}^{n} v(s_i, t_i)$$

Figure 3: All references sentence error rate formula

Other scoring metrics are:

- *Position independent error rate* (PER): this measure is similar to the *word error rate*, but it ignores or does not take into account the positions of the words in the reference sentences. It is a more suitable metric to evaluate the system for tasks where the source and target language words are arranged in different ways and the output sentence admits different rearrangements.
- *BiLingual Evaluation Understudy* (BLEU): this measures the translation closeness between a candidate translation and a set of reference translations with a numerical metric. It scores the ngram precision (unigrams, bigrams, trigrams and 4-grams) with regard to a sample of reference translations. An N-gram is a subsequence of n-elements from such a sequence (Lin and Hovy 2003).
- *National Institute of Standards and Technology* (NIST): this algorithm is based on the BLEU metric, but with some modifications. Where BLEU simply calculates n-gram precision, adding equal weight to each one, NIST also calculates how informative a particular n-gram is. That is, n-gram precisions are weighted by the n-gram frequencies, to place more emphasis on the less frequent (and more informative) n-grams. Another significant difference with respect to BLEU is that NIST computes the arithmetic mean of the n-gram precision, also with a length penalty (Goutte 2006).

Up to this point, comparison of sentences has been discussed without evaluating which kind of error or discrepancy occurred between two sentences. However, there are other systems that automatically evaluate translation assessments made by subjective or human methods (Wikipedia 2008). Amongst these are:

- *Subjective sentence error rate* (sSER): in this measure the score ranges from 0 to 1, signifying from a perfect translation to a nonsense sentence. Originally, Nießen (2000) proposed a score scale from 0 to 10 but has since concluded that scale is too wide and that six or seven quality classes it would function better.
- *Information item error rate* (IER): for this measure, the sentence is segmented into information items. A person assesses whether the information from each item is found in the translation. Hence, it can be checked whether, in a wrong translation of the sentence, there are parts that are correctly translated (Nießen et al. 2000).

Our aim is to customize the system of automatic-translation evaluation *subjective sentence error rate* with our research from German to Spanish and from French to

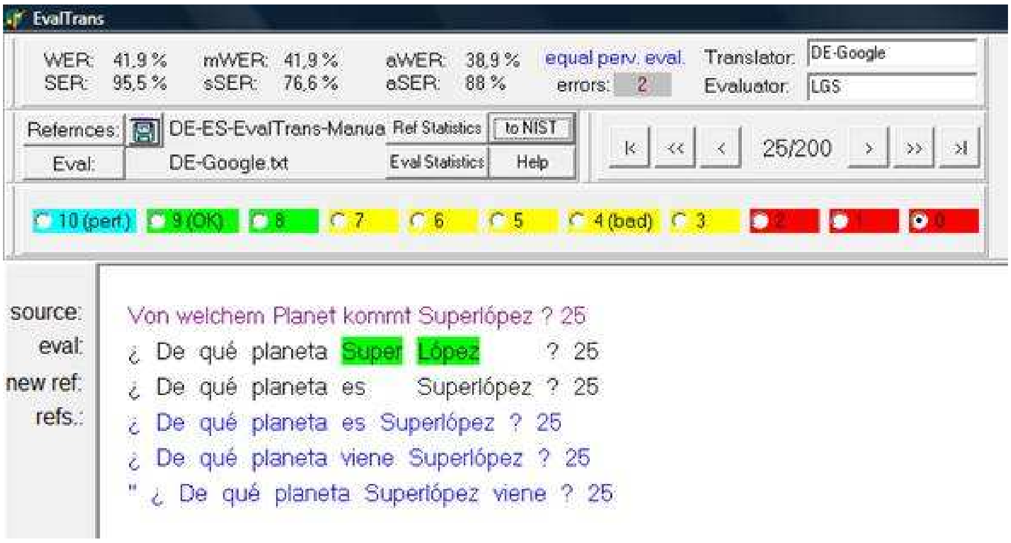
Spanish. In addition, we propose an evaluation form adapted specifically to the question-answering systems. A *perfect* translation is not the aim of this study, but, rather, a translation which is able to preserve the characteristics of the questions and consequently enable the system to locate the suitable answers.

Analysis with EvalTrans

The evaluation process was carried out using EvalTrans software (Nießen *et al.* 2000) in its graphic version designed for use with Windows (Tomás *et al.* 2003). This tool can be used freely online, for the evaluation of automatic translation. It provides us with statistics, such as the average Levensthein distance standardized to the length of the target sentence and calculates previous rates by means of automatic metrics (Dabbadie *et al.* 2002).

One advantage of this program is that it can compare the results of each translator with the other translators studied for the same set of phrases. It works with the follow indicators: WER, mWER, aWER, SER, mSER and sSER (described above). These measures are widely used and they consider important the order of words in the sentence. In addition, it allows for a qualification in the assessment of the translation with a manual and subjective supervision. It is intended to correct the failings of an automatic system without a lexicon when seeking to detect several possible ideal sentences (it supports only a single perfect sentence for comparison). It also includes grading for the assessment (Figure 4).

The selected translators were evaluated independently (e.g., *word error rate* (WER)) and combined (e.g., *all references word error rate* (aWER) and *all references word sentence rate* (aSER)) . EvalTrans highlights the differences in the comparison of the two sentences, which is evaluated with respect to the reference.



WER: 41,9 %   mWER: 41,9 %   aWER: 38,9 %   equal perv. eval.   Translator: DE-Google
SER: 95,5 %   sSER: 76,6 %   aSER: 88 %   errors: 2   Evaluator: LGS

Refernces: DE-ES-EvalTrans-Manua
Eval: DE-Google.txt
25/200

○ 10 (perf.) ○ 9 (OK) ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 (bad) ○ 3 ○ 2 ○ 1 ● 0

source: Von welchem Planet kommt Superlópez ? 25
eval: ¿ De qué planeta Super López ? 25
new ref: ¿ De qué planeta es Superlópez ? 25
refs.: ¿ De qué planeta es Superlópez ? 25
¿ De qué planeta viene Superlópez ? 25
" ¿ De qué planeta Superlópez viene ? 25

Figure 4: Example of EvalTrans

The metrics used are designed for phrases and questions are short phrases with a predictable structure according to the grammar of the language.

EvalTrans is a tool for evaluating translations using metrics, which are considered as important as the word order. In an effort to mitigate the rigidity of some of these, the *all references word error raqte* (aWER) and the *all references sentence error rate* (aSER) are calculated. The latter takes into account more than one alternative when the sentence assessed is compared or checked with the reference or model-phrase collection. This metric evaluation is made fully automatic by calculating the *subjective sentence error rate* (sSER) with human values (or subjective ratings rather than the match or not match automatic methods), which considers the functionality of the translation in a question-answering system. The increase in the number of reference phrases in human evaluation also changes the rates of aWER and aSER.

Analysis of results

The results shown are averages, based on the yield of the online translators, given the values obtained on applying the measures described above and the values that resulted from the human assessment of each question translated.

Effectiveness of the online translators according to indicators used

Tables 1 and 2 show the values based on the set of 200 questions in terms of *word error rate* (WER) and *sentence error rate* (SER) for the *Google Translator*, *Promt* and *Worldlingo* in automatic evaluations, from German and French into Spanish.

GERMAN	Google	WorldLingo	Promt
WER	41.9%	57.6%	54.4%
SER	95.0%	98.5%	99.0%

Table 1: Automatic evaluation of the translations from German to Spanish

FRENCH	Google	WorldLingo	Promt
WER	43.2%	40.8%	39.6%
SER	95.5%	93.0%	90.0%

Table 2: Automatic evaluation of the translations from French to Spanish

The high values found for the *sentence error rate* (SER) can be attributed to the need for the evaluation software to find a string from the online translator that is identical (with the same words and in the same order) to the reference phrase in order to calculate the edition rate. Any variation, even a minor one, is interpreted as an erroneous phrase (in our case an erroneous question) and is left out. Tables 3 and 4 below reflect that the *subjective sentence error rate* (sSER) measure aims for greater precision than the SER measure, by taking into account human evaluation and the corresponding acceptance or rejection of the phrase supplied by the online translator, judged as correct or incorrect. Our objective, however, is merely to identify the translating program that generates the most functional input for a translingual question-answering program.

In addition, the coefficients corresponding to the *all references word error rate*, the *sentence error rate* and *all references sentence error rate* do indeed vary in conjunction with human intervention (see Tables 3 and 4). For instance, the *subjective sentence error rate* measure takes the scores for each one of the phrases already translated and evaluated. The *all references word error rate* measure, meanwhile, gathers all the reference phrases that have been considered subsequently as such by a human translator. These tend to be proposed by the human evaluator as *new reference* after the reduction of the edit distance; or else, a candidate phrase is scored with a maximum mark. The evaluating program adopts the phrase of reference that is most similar to the group of phrases of reference already existing, not just the first sentence of reference included *a priori*.

The ranking of the online translation programs analysed with regard to their effectiveness means that the best translator of the three would be the one scoring the lowest index (lowest occurrence of errors), especially evident with *subjective sentence error rate* and *all references word error rate*, which account for human assessment.

Because the applied measures for automatic evaluation do not make a comprehensive syntactic analysis (noting the position of the words in the phrase), the error indexes are greater in German. The fact that the edit distance registers not only the existence of words in the sentence, but also their position, leads to higher error rates when the German language is involved, since alteration in the order of the elements is identified as an error (Tillman *et al.* 1997).

Accordingly, grammatical similarities between the French and Spanish languages result in fewer errors (see Tables 1 and 2). It bears noting that only in the case of *Google Translator* were the error indexes in conjunction with words (*word error rate*) higher for French than for German. After the subjective assessments of the translations by the application of *all references word error rate* (aWER), *subjective sentence error rate* (sSER) and *all references sentence error rate* (aSER), the results were as follows:

GERMAN	Google	WorldLingo	Promt
aWER	57.6%	54.6%	50.4%
sSER	90.2%	91.3%	77%
aSER	88%	94%	91%

Table 3: Indicators calculated with human assessment of the translations from German to Spanish

FRENCH	Google	WorldLingo	Promt
aWER	36.7%	29.7%	27.5%
sSER	70.4%	53.7%	55.5%
aSER	87.5%	78.5%	75%

Table 4: Indicators calculated with human assessment of the translations fromFrench to Spanish

Practically all the values decreased with human assessment, meaning that the error indexes were reduced. The consideration of various alternatives as acceptable leads to a greater yield of questions of reference for calculating *all references word error rate* and *all references sentence error rate*. The graphic display of the assessment values found with regard to the translations into Spanish from German (Figure 5) and from the French (Figure 6) for the automatic translators evaluated reveal that the error indexes per word (WER and aWER) were lower when the source language was French.

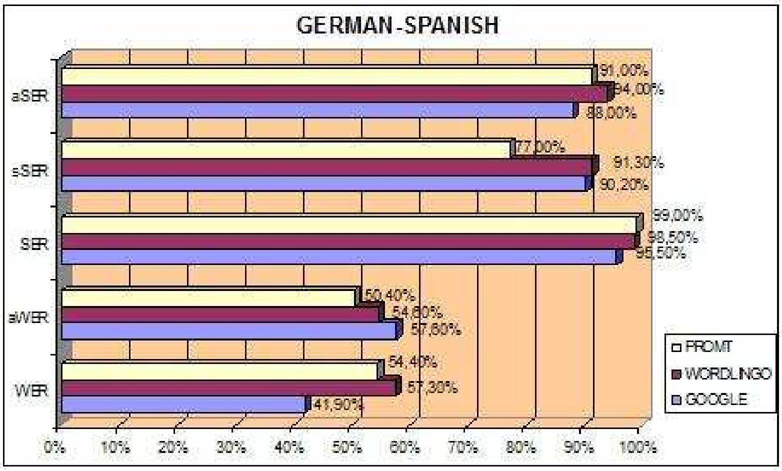


Figure 5 : Comparison of overall values after human assessment (German-Spanish)

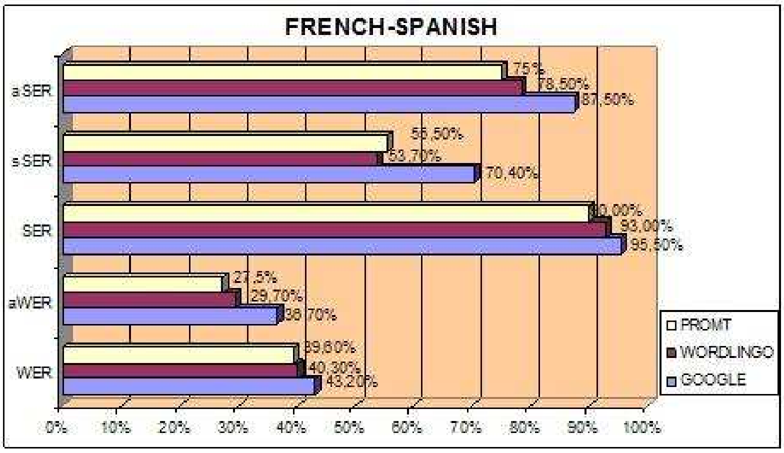


Figure 6: Comparison of values after human assessment (French-Spanish)

Likewise, the percentages derived from the errors per phrase, whether subjective (the *subjective sentence error rate* (sSER)) or automatic (the *sentence error rate* (SER) and the *all references sentence error rate* (aSER)), were lower for the translation from French than from German. One reason would be that these measures do not register any coincidence of words if the automatic translator has altered the word order with regard to the phrase of reference.

We may also interpret the *word error rate* indicators depending on the kind of sentence or phrase. In the 200 phrases studied for each language, regardless of their length, it would be necessary to change, reorder, replace, or modify between four and six words of the string in order to achieve the ideal translation or equal the reference translation (Table 5).

Edit Distance	Google	WorldLingo	Promt
GERMAN	4.29	5.85	5.57
FRENCH	4.42	4.12	4.05

Table 5: Average number of words that must be modified to arrive at the "perfect" translation

Evaluation of the error rate without taking into account the order of the terms within the phrase is reflected the *subjective sentence error rate* measure: when translating from German into Spanish, this error rate was greater than for French translated into Spanish. In other words, evaluations of the translations are better when the word order is overlooked. For French (Table 4), where grammatical similarity leads to a considerable reduction in error, the *Promt* translator offers the best results, followed by *Worldlingo*. However, when the phrases are translated from the German language, Google provides the best results in the automatic measurements (Table 3). The values are lower (better) for *Promt* than for *Google*, with *Worldlingo* taking last place. The fact that this measure relies on human assessment bears weight in the selection of the ideal online translator, since the main criterion to be upheld is the effectiveness of the resulting translation as input in a question-answering system.

Human evaluation of the translations

For the manual evaluation of the translations generated by automatic online translators, we applied the Likert scale (Uebersax 2006), using six levels. We were thus able to reduce the excessive range of the classification method of Niessen *et al*. (2000), with its eleven levels, adopted by EvalTrans. Taking into account the finality of the translation, the assessment implied that errors such as the position of the elements in the string would not have to be penalized to the same degree as ambiguity, or the loss of some characteristic of the question (e.g., interrogative adverb, or the entity to which the question refers). The resulting assessments were zero when the phrase was totally incorrect and meaningless as a translation, and therefore useless as a question for input in a question-answering system; up to five when the phrase was considered *perfect*. These values were then used to calculate the indexes given below.

The automatic evaluation compares the output with the reference sentence provided by the Cross-Language Evaluation Forum. First, we have a sample of 200 questions in Spanish and their French and German translations. The human evaluation is made without taking into account reference sentences to compare with the translator's output. The role of a professional human translator is to monitor the outputs of the three new online translators and rating on a scale of 0 to 5 if the new translation is suitable from a functional or operational standpoint and without relying on one or more reference phrases. On the other hand, these measurements were complemented automatically with manual evaluation, in which a professional human translator who works in these three languages, has assigned scores (from 0 to 5). to assess the quality of the translation of each one of the phrases in both language pairs. Also, it was possible to identify the different types of errors that occurred throughout the translation process. The manual evaluation was consistently higher than the phrases translated into Spanish and we should bear in mind that a functional criterion was taken into account in this study to determine the translation quality. Figures 5 and 6 show the results of this human assessment.

As discussed in the section above, according to the *subjective sentence error rate* index, in the case of German, *Promt* (77%) is the best translator, followed by *Google* (with 90.2%). By contrast, when dealing with French, *Worldlingo* (53.7%) proved to be the best translator, although *Promt* (55.5%) had a very similar rating. According to the human assessment, the behaviour of the three online translators analysed is more irregular (presenting a greater standard deviation) in the case of the translations from French than for those from German (Table 6). The resulting values were lower for German but more homogeneous for French.

	Google	WorldLingo	Promt
GERMAN-SPANISH	1.820	1.573	1.522
FRENCH-SPANISH	2.010	2.132	2.136

Table 5: Standard deviation calculated over human assessments

Figure 6 shows the Bell curve and the normal distribution of the values found. Nevertheless, the irregularity of the translations from French is reflected in greater differences between the translators. In this sense, it is notable that *Worldlingo*, when translating from French (Figure 7), scores higher in the evaluation of its translations, with a mean value of 2: a nearly correct and quite consistent translation. This is seen in the asymmetric distribution tending toward the left, or lower values on the scale used. The normal distribution is almost fully centred, as in the case of the *Google Translator*; yet its values are much lower (Figures 7 and 8).

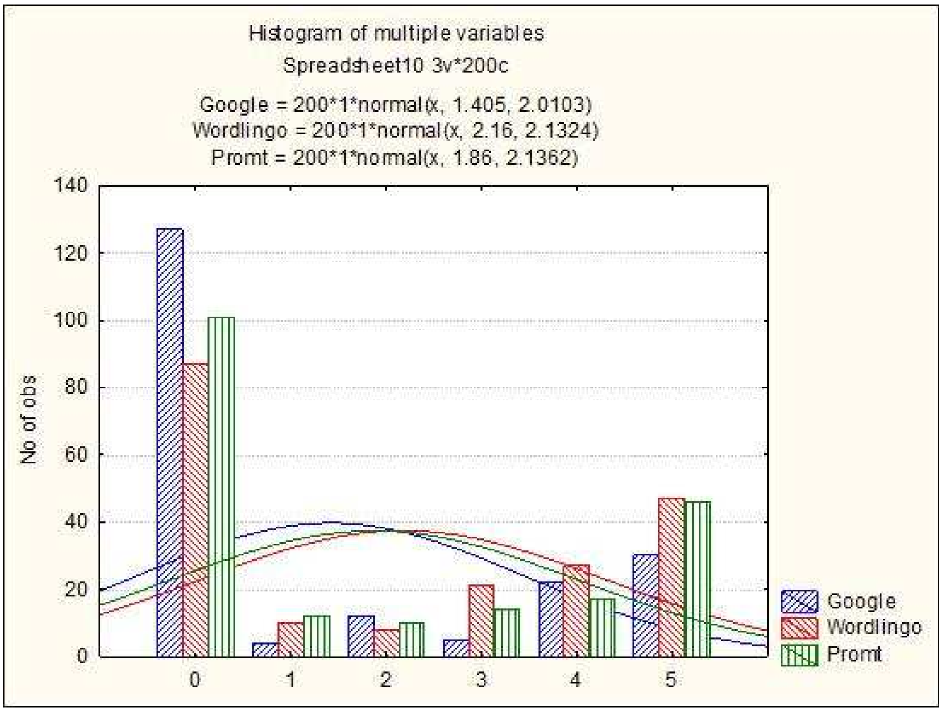


Figure 7: Distribution of the evaluations involving French

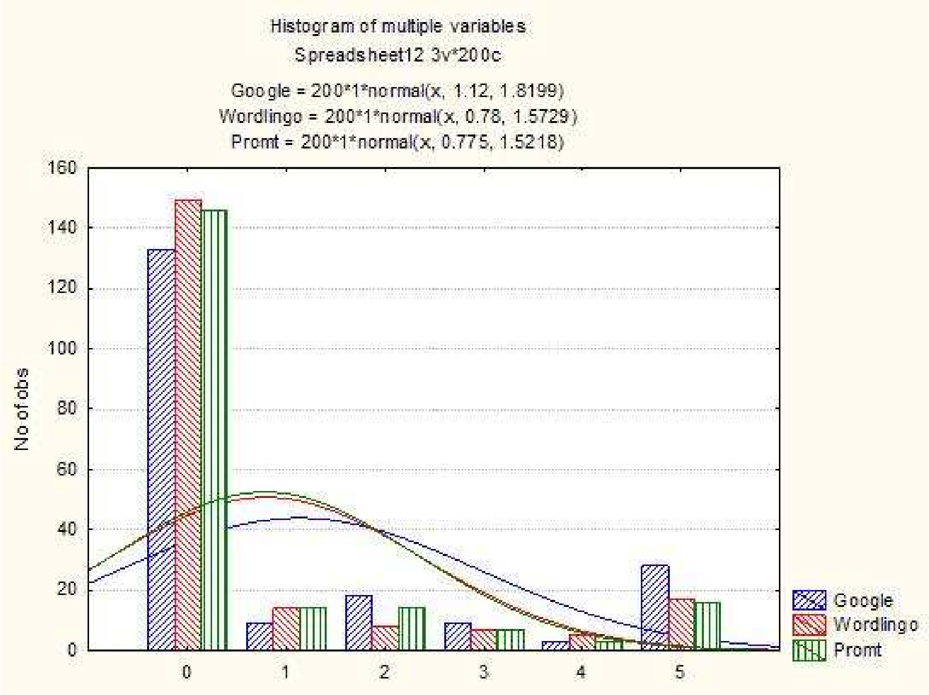


Figure 8: Distribution of the evaluations involving German

Conclusions

We studied the automatic online translators *Google Translator, Promt* and

Worldlingo, applying different means of evaluation. Strictly automatic evaluation (in the absence of subjective assessment) was found to render high error indexes that are scarcely representative.

Automatic translators are the most widely used resource for multilingual question-answering systems, followed by textual corpora and dictionaries (Nguyen *et al.* 2009). Despite the problems of ambiguity and non-optimal quality of the texts (especially in case of restricted domain systems), automatic translators turn out to be one of the most economical and simple tools to integrate with these systems. However, while this is still the favorite tool of choice for developers, individually or in combination with other language resources, a tendency can be discerned in recent years to incorporate other resources. Indeed the language corpora are also useful, especially for question-answering multilingual systems in a specialized domain, since they are made by professional human translators.

The textual corpora that are located on the Web with pages available in several languages can overcome the problems of computational cost and storage, very common in this type of resource. Furthermore, the grammar and ambiguity problems have decreased the use of dictionaries as a resource for translation in question-answering systems; however, there is increased use of new resources such as ontologies, Wikipedia, databases, thesauri, multilingual lexico-semantic networks, such as EuroWordNet, or tools such as computational grammars. Sometimes, a single application or tool is not complete enough to solve the problem of cross-lingual communication so that the use of several could offer better results.

The error indexes from automatic evaluation are higher when the translation is from the German into Spanish, because the most frequently used measurements for evaluating translations use indicators that compare word-by-word, looking for the very same order of elements in the translation produced online as in the initial reference phrase. Therefore, the syntactic errors detected are more numerous because of basic grammatical differences between the source language (German) and the target language (Spanish). Also greater are the error indexes resulting from subjective assessment, as made here, owing to the capacities of the tools themselves in translating from German into Spanish. The grammatical similarities between French and Spanish tend to give a lower error index.

Our findings identify *Promt* as the most reliable translator for both pairs of languages overall. However, an even more reliable option would be to use two different translators, depending on which of the two source languages is being used. In this case, the selected translators would be *Google Translator* or *Promt* for the German language and *Promt* for French.

At this point, evaluation measures for automatic translators need to be explored in greater depth in order to arrive at means that provide more flexible criteria (not strictly the dichotomy of exact match vs. not exact match) for assessing the translated phrase to be evaluated. This may be achieved using a larger set of lexical strings of reference, or by adjusting better to the order of the elements within the phrase, or even through some consideration of the roots and canonical forms, as well as the words with their complete flexive and identifying traits (exact match) . Improving the units of measure and techniques for automatic evaluation constitutes a research front which parallels the improvement of automatic translating systems themselves. Moreover, it would prove beneficial if the different tools now being used or developed for the evaluation of translations, such as EvalTrans, and the various research studies undertaken were to use the same scale of human assessment. This would make it easier to introduce data and to quantify the measures applied by human assessments (such as the *subjective sentence error rate* measure).

The translation mistakes of the interrogative particles identified in Table 2 vary according to the question and the online automatic translator used. The findings show that there is no direct relationship between the type of interrogative particle and the resulting error. It is worth noting, however, that the mismatch among automatic translators when they make an error indicates that the information stored in each online automatic translator could complement each another so as to improve their efficiency.

Grammar can have an impact on meaning. For an instance in English, an apostrophe makes the contraction *it's*, which means *it is*, while the same three letters spell *its*, the possessive third person pronoun or adjective for *it*. The absence or presence of the apostrophe is not an option and changes the meaning of the word entirely and, therefore, an online translator should know the difference. In the same sense, a good automatic translator needs the set of rules that apply to the correct usage of a language. In our research, we have discovered that polysemous verbs in the target language lead to the wrong translation. This is the case of the verb *sein* in German or *être* in French (*to be* in English) which has two possible meanings in Spanish, *ser* or *estar*. This polysemous situation could be solved with a syntactic analysis of the sentence. In this example, we could introduce a grammatical rule explaining that the verbs *sein* or *être* can be translated as *estar* if there is any noun of location in the sentence.

Finally, it bears mentioning here that although the quality of automatic online translations may be poor, the demand for this type of translation tool is widespread and growing, especially in the multilingual context of the Internet. Therefore, such services will be demanded to an increasing extent in the future and we should concentrate continuous effort on their improvement. In future studies our research team will follow this line deeper into the design of efficient and effective multilingual question-answering systems.

Acknowledgements

We wish to thank the anonymous referees for their helpful comments.

About the authors

Lola García-Santiago is a Doctor of Documentation, University of Granada, where she works as Professor. At present she teaches at both the University School of Documentation and the School of Translation and Interpretation. She is also part of the associated unit of the SCImago research group, in Spain’s Higher Council for Scientific Research (CSIC), where her lines of work and published articles focus mainly on Web-mining, Internet information retrieval, question-answering systems and gray literature. She can be contacted at mdolo@ugr.es

María-Dolores Olvera-Lobo holds a Ph.D. in Documentation and is full Professor in the Department of Communication and Documentation of the University of Granada and also teaches in the School of Translation and Interpretation. As member of the associated unit of the SCImago team, she participates in a number of research projects now under way and has authored or co-authored several books and numerous articles published in specialized journal. She can be contacted at molvera@ugr.es

References

- Abusalah, M.,Tait, J. & Oakes, M. (2005). Literature review of cross language information retrieval. *Proceedings of World Academy of Science, Engineering and Technology*, **4**, 175-177.
- Belkin, N.J. & Croft, W. B. (1987). Retrieval techniques. *Annual Review of Information Science and Technology*. **22** 109-146.
- Bloom, B.S. & Krathwohl, D.R. (1956). Taxonomy of educational objectives: the classification of educational goals, by a committee of college and university examiners. Handbook 1: Cognitive domain. New York, NY: Longmans.
- Bos, J. & Nissim, M. (2006). Cross-lingual question answering by answer translation. In C. Peters (Ed.), *Working Notes for the CLEF 2006 Workshop, 20-22 September, Alicante, Spain*. Retrieved 17 November, 2010 from http://www.clef-campaign.org/2006/working_notes/workingnotes2006/bosCLEF2006.pdf (Archived by WebCite® at http://www.webcitation.org/5uJHOlGWR)
- Chang Y., Xu H.B. & Bai, S. (2003). TREC 2003 question answering track at CAS-ICT. In . M. Voorhees and Lori P. Buckland (Eds.). *The Twelfth Text Retrieval Conference (TREC 2003), Gaithersburg, Maryland*. (pp. 147-151). Retrieved 17 November, 2010 from http://trec.nist.gov/pubs/trec12/papers/chinese-acad-sci.qa.final.pdf (Archived by WebCite® at http://www.webcitation.org/5uJIJ2aEx)
- Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T., Jensen, J. & Hersh, W. (2006). The CLEF 2005 cross-language image retrieval track. In C. Peters, F.C. Gey, J. Gonzalo, H. Müller, G.J.F. Jones, M. Kluck, *et al.* (Eds.), *Accessing Multilingual Information Repositories. 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005*, , (pp. 535-557). Berlin, Heidelberg: Springer-Verlag. (Lecture Notes in Computer Science, 4022)
- Clough, P., Müller, H. & Sanderson, M. (2005), (2004). The CLEF cross-language image retrieval track (ImageCLEF) 2004. In P. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton & A.W.M. Smeulders, (Eds.). *Proceedings of the third International Conference Image and Video Retrieval (CIVR) 2004, Dublin, Ireland, July 21-23, 2004*. (p. 2066.) Berlin, Heidelberg: Springer-Verlag. (Lecture Notes in Computer Science, 3115)
- Cross-Language Evaluation Forum. (2008). Guidelines for the participants in QA@CLEF 2008. Retrieved 19 November 2010 from http://clef-qa.fbk.eu/2008/download/QA@CLEF08_Guidelines-for-Participants_new.pdf (Archived by WebCite® at http://www.webcitation.org/5utnRpnH0)
- Dabbadie, M., Hartley A., King M., Miller K.J., Mustafa El Hadi W., Popescu-Belis A. *et al.* (2002). A hands-on study of the reliability and coherence of evaluation metrics. In M. King, (Ed.). *Workbook of the LREC 2002 Workshop on Machine Translation Evaluation: Human Evaluators Meet Automated Metrics. Las Palmas de Gran Canaria, Spain, 27 May 2002*, (pp.8-16). Retrieved 18 november 2010 http://www.mt-archive.info/LREC-2002-WS-MTEval.pdf (Archived by WebCite® at http://www.webcitation.org/5utoE1adF)
- Forehand, M. . (2005). Bloom's taxonomy: original and revised. In M.

How to cite this paper

García-Santiago, L & Olvera-Lobo, M. D.(2009). "Analysis of automatic translation of questions for question-answering systems" *Information Research*, 15(4) paper 450. [Available at http://InformationR.net/ir/15-4/paper450.html]

Find other papers on this subject

Scholar Search | Google Search | Windows Live

Check for citations, using Google Scholar

Bookmark This Page

1103

Contents | Author index | Subject index | Search | Home